

Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds

Jiannan Lu, Peng Ding* and Tirthankar Dasgupta†

Abstract

Under the potential outcomes framework, causal effects are defined as comparisons between the treatment and control potential outcomes. Unfortunately, however, the average causal effect, often the parameter of interest, is generally not well defined for ordinal outcomes. To address this problem, we propose to use two causal parameters that are defined as the probabilities that the treatment is beneficial and strictly beneficial for the experimental units. These two causal parameters are well defined for any outcomes and of particular interest for ordinal outcomes. These parameters, though of scientific importance and interest, depend on the association between the potential outcomes and are therefore, without further assumptions, not identifiable from the observed data. In this paper, for ordinal outcomes we derive the sharp bounds of the two causal parameters using only the marginal distributions, without imposing any assumptions on the joint distribution of the potential outcomes. Because we define the causal effects and derive the bounds based on the potential outcomes themselves, the theoretical results can be incorporated into any models of the potential outcomes, and are applicable to randomized experiments, unconfounded observational studies, and randomized experiments with noncompliance. We illustrate our methodology via numerical examples and real-life applications.

Keywords: Linear programming; Monotonicity; Noncompliance; Partial identification; Potential outcome; Stochastic dominance.

*Corresponding author

†Jiannan Lu is Data Scientist (E-mail: jiannl@microsoft.com), Analysis and Experimentation, Microsoft Corporation, Redmond, WA 98052, U.S.A. Peng Ding is Assistant Professor (E-mail: pengdingpku@berkeley.edu), Department of Statistics, University California, Berkeley, CA 94270, U.S.A. Tirthankar Dasgupta is Associate Professor (E-mail: dasgupta@stat.harvard.edu), Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

1. Introduction

The potential outcomes framework (Neyman 1923; Rubin 1974) permits defining causal effects as comparisons between the potential outcomes under treatment and control. The average causal effect, generally the parameter of interest ever since the seminal work of Neyman (1923), is not applicable to ordinal outcomes, because average outcomes themselves are not well defined. Ordinal outcomes are common in applied research, and the generalized linear model literature (cf. Agresti 2010) has discussed ordinal outcomes extensively. Unfortunately, however, although the model parameters of the generalized linear models are useful summaries of the data, they are often not direct measures of the causal effects of interest (Freedman 2008). Moreover, statistical inference often requires correctly-specified models, and when the generalized linear model assumptions are violated, the interpretations of the parameters often become obscure. The causal inference literature mainly focuses on the average causal effect, and does not give a thorough investigation of ordinal outcomes. Rosenbaum (2001) discussed causal inference for ordinal outcomes under the monotonicity assumption that the treatment is beneficial for all units. Cheng (2009) and Agresti (2010) discussed various causal parameters under the assumption of independent potential outcomes. Volfovsky et al. (2015) exploited a Bayesian strategy, requiring a full parametric model on the joint values of the potential outcomes. Díaz et al. (2016) proposed to use a causal parameter that did not rely on the assumption of the proportional odds model for ordinal outcomes.

For ordinal outcomes, we propose to use two causal parameters measuring the probabilities that the treatment is beneficial and strictly beneficial for the experimental units, which play important roles in decision and policy making for randomized evaluations with ordinal outcomes. Because these two causal parameters depend on the association between the treatment and control potential outcomes, they are generally not identifiable from the observed data. Without imposing any assumptions about the underlying distributions of, or the association between, the potential outcomes, we sharply bound them by using the marginal distributions of the potential outcomes. Mathematically, deriving the sharp bounds of the proposed causal parameters is closely related to a classical probability problem posed by A. N. Kolmogorov (c.f. Nelsen 2006), which is a non-trivial task for ordinal outcomes. We believe this is a major contribution to the literature.

Because these bounds hold for causal parameters defined by the potential outcomes themselves,

they hold without any modeling assumptions, and therefore can be incorporated flexibly into any chosen models of the potential outcomes in practice. Furthermore, they are directly applicable to randomized experiments, unconfounded observational studies, and randomized experiments with noncompliance. In randomized experiments, we can identify the bounds immediately, and additionally, sharpen the bounds by exploiting covariate information under certain modeling assumptions. In observational studies, if the treatment assignment is unconfounded given the observed covariates, we can identify the bounds by propensity score weighting (Rosenbaum and Rubin 1983; Hirano et al. 2003). Furthermore, we extend the theory to accommodate noncompliance, because it often arises in practical randomized evaluations.

The paper proceeds as follows. Section 2 introduces the potential outcomes framework for causal inference for ordinal outcomes, and proposes two causal parameters that are natural measures of causal effects and are of practical importance. Section 3 derives the sharp bounds of the proposed causal parameters. Section 4 generalizes the bounds to noncompliance. Section 5 discusses statistical inference of the bounds. Sections 6 and 7 present numerical and real examples to illustrate the theoretical results. We conclude in Section 8, prove the main theorem in the Appendix, and relegate other technical details to the Supplementary Material.

2. Causal Inference for Ordinal Outcomes

2.1. Potential Outcomes

We consider a study with N units, a binary treatment, and an ordinal outcome with J categories labeled as $0, \dots, J-1$, where 0 and $J-1$ represent the worst and best categories. Under the Stable Unit Treatment Value Assumption (Rubin 1980) that there is only one version of the treatment and no interference among the units, we define the pair $\{Y_i(1), Y_i(0)\}$ as the potential outcomes of the i th unit under treatment and control, respectively. Let

$$p_{kl} = \text{pr} \{Y_i(1) = k, Y_i(0) = l\} \quad (k, l = 0, \dots, J-1)$$

denote the proportion or probability of units whose potential outcome is k under treatment and l under control. The probability notation “ $\text{pr}(\cdot)$ ” is either for a finite population of N units or for a

super population, depending on the question of interest. The probability matrix $\mathbf{P} = (p_{kl})_{0 \leq k, l \leq J-1}$ summarizes the joint distribution of the potential outcomes. We denote the row and column sums of \mathbf{P} by

$$p_{k+} = \sum_{l'=0}^{J-1} p_{kl'}, \quad p_{+l} = \sum_{k'=0}^{J-1} p_{k'l} \quad (k, l = 0, 1, \dots, J-1).$$

The vectors $\mathbf{p}_1 = (p_{0+}, \dots, p_{J-1,+})^T$ and $\mathbf{p}_0 = (p_{+0}, \dots, p_{+,J-1})^T$ characterize the marginal distributions of the potential outcomes under treatment and control, respectively.

2.2. Causal Parameters for Ordinal Outcomes

We discuss the existing causal parameters for ordinal outcomes, and the motivation behind proposing new ones. Any causal parameter is a function of the probability matrix \mathbf{P} . Unfortunately, the average causal effect is not well defined for ordinal outcomes. Instead, we can use the distributional causal effects (cf. Ju and Geng 2010)

$$\Delta_j = \text{pr}\{Y_i(1) \geq j\} - \text{pr}\{Y_i(0) \geq j\} = \sum_{k \geq j} p_{k+} - \sum_{l \geq j} p_{+l} \quad (j = 0, \dots, J-1) \quad (1)$$

to measure the difference between the marginal distributions of potential outcomes at different levels of j . However, unless the distributional causal effects Δ_j 's have the same sign for all j , it is difficult to decide whether the treatment or the control is preferable. We may use $\sum_{j=1}^{J-1} \omega_j \Delta_j$ to measure the treatment effect, but such a measure depends crucially on the weights ω_j 's. We illustrate this point by using the following numerical example.

Example 1. Let $\mathbf{p}_1 = (1/5, 3/5, 1/5)^T$ and $\mathbf{p}_0 = (2/5, 1/5, 2/5)^T$, with $\Delta_0 = 0$, $\Delta_1 = 1/5$ and $\Delta_2 = -1/5$. The treatment is beneficial at level 1, but not at level 2. In this case, distributional causal effects do not provide straightforward guidance for decision making.

When $\Delta_j \geq 0$ for all j , $Y(1)$ stochastically dominates $Y(0)$. When this pattern appears in real data applications, practitioners often fit a proportional odds model (Agresti 2010) and summarize the overall effectiveness of the treatment by a single odds ratio parameter. Although such summary parameter may be useful in certain cases, its causal interpretation is unclear. Moreover, when the data does not present the stochastic dominance pattern as in Example 1, summarizing the treatment effect by the single odds ratio parameter of a wrong model often gives misleading conclusions.

Volfovsky et al. (2015) studied the conditional medians

$$m_j = \text{med} \{Y_i(1) \mid Y_i(0) = j\} \quad (j = 0, \dots, J-1), \quad (2)$$

which is a set containing all values of k such that $\sum_{k'=0}^k p_{k'j} \geq p_{+j}/2$ and $\sum_{k'=k}^{J-1} p_{k'j} \geq p_{+j}/2$. By definition, the conditional medians may not be unique, and they are only well defined for j with $p_{+j} > 0$. Moreover, they are not direct measures of the treatment effect itself.

We propose to use two causal parameters that measure the probabilities that the treatment is beneficial and strictly beneficial for the experimental units:

$$\tau = \text{pr} \{Y_i(1) \geq Y_i(0)\} = \sum_{k \geq l} p_{kl}, \quad \eta = \text{pr} \{Y_i(1) > Y_i(0)\} = \sum_{k > l} p_{kl}. \quad (3)$$

The causal parameters τ and η are measures of causal effects that are well defined for any types of outcomes, and of particular interest to ordinal outcomes. Similar causal measures appeared in biomedical (Gadbury and Iyer 2000; Newcombe 2006a,b; Zhou 2008; Huang et al. 2015; Demidenko 2016) and social sciences (Heckman et al. 1997; Djebbari and Smith 2008; Fan and Park 2010; Fan et al. 2014). In practice, we suggest using the pair (τ, η) as measures of causal effects on ordinal outcomes. For example, if the sharp null holds, i.e., $Y_i(1) = Y_i(0)$ for all units i , then $\tau = 1$ and $\eta = 0$. In this case, using only τ may be misleading. Nevertheless, we argue that the parameter τ is as important as η . Because $1 - \tau = \text{pr} \{Y_i(0) > Y_i(1)\}$, the value of τ determines the probability that the control is strictly beneficial for the experimental units. Due to the symmetry of treatment and control labels, τ and η are equally useful for real data analysis.

We use the following numerical example to show the values of m_j , τ and η .

Example 2. Consider the following probability matrix:

$$\mathbf{P} = \begin{pmatrix} 0 & 1/6 & 1/6 \\ 0 & 1/6 & 0 \\ 0 & 1/3 & 1/6 \end{pmatrix}.$$

In this case, m_0 is not well defined, m_1 is 1, and $m_2 = \{0, 1, 2\}$. However, we have $\tau = 2/3$ and $\eta = 1/3$, i.e., two thirds of the population benefit from the treatment and one third strictly benefit.

The causal parameters τ and η in (3) are well defined for both finite populations and super populations. They are functions of the potential outcomes, which distinguishes them from the parameters in super population models. When the models are mis-specified, the interpretations of the corresponding model parameters are often obscure. We have already discussed this issue for the proportional odds model. Our causal parameters τ and η are closely related to the relative treatment effect $\alpha = \text{pr}\{Y_i(1) > Y_i(0)\} - \text{pr}\{Y_i(1) < Y_i(0)\}$ previously studied under the assumption of independent potential outcomes (Agresti 2010). This relative treatment effect α and the causal parameters we proposed have a simple algebraic relationship, i.e., $\alpha = \tau + \eta - 1$. Therefore, our newly proposed causal parameters τ and η determine α . Furthermore, these causal parameters are also related to the notation of “probability of causation” (Pearl 2009), because their direct interpretations are the probabilities or proportions that the treatment affects the outcome on the individual level. It is for these reasons that we advocate using τ and η as causal effect measures for ordinal outcomes.

3. Sharp Bounds on the Proposed Causal Estimands for Ordinal Outcomes

3.1. Closed-Form Expressions of Sharp Bounds

The definitions of τ and η involve the association between the treatment and control potential outcomes. Because we can never jointly measure the potential outcomes, the observed data do not provide full information about their association, rendering the causal parameters τ and η not identifiable. To partially circumvent this difficulty, we focus on the sharp bounds of τ and η , which are the minimal and maximal values of τ and η under the constraints of the following marginal distributions:

$$\sum_{l'=0}^{J-1} p_{kl'} = p_{k+}, \quad \sum_{k'=0}^{J-1} p_{k'l} = p_{+l}, \quad p_{kl} \geq 0 \quad (k, l = 0, \dots, J-1). \quad (4)$$

The sharp bounds depend only on the marginal distributions of the potential outcomes. Deriving the sharp bounds is equivalent to solving linear programming problems, because the objective functions in (3) and the constraints in (4) are all linear. Previous literature (Huang et al. 2015) used a numerical method to solve the linear programming problem for η . Fortunately, we can derive

closed-form solutions of the above linear programming problems for both τ and η .

In this paper, we not only derive the sharp bounds of the causal parameters of interest, but also construct explicitly the probability matrices that attain these bounds. First we state a theorem on the sharp bounds of τ , which is the foundation for the remaining theorems and corollaries.

Theorem 1. The sharp lower and upper bound of τ are

$$\tau_L = \max_{0 \leq j \leq J-1} (p_{+j} + \Delta_j), \quad \tau_U = 1 + \min_{0 \leq j \leq J-1} \Delta_j.$$

The bounds in Theorem 1 are closely related to the distributional causal effects in (1), and we can interpret them as the conservative and optimistic estimates of the probability that the treatment is beneficial to the outcome. Furthermore, the following corollary demonstrates that the sharp upper bound τ_U is related to the stochastic dominance assumption, i.e., $\Delta_j \geq 0$ for all j .

Corollary 1. The causal parameter $\tau_U = 1$, if and only if the marginal probabilities \mathbf{p}_1 and \mathbf{p}_0 satisfy the stochastic dominance assumption.

The above corollary implies that for any marginal probabilities satisfying the stochastic dominance assumption, there exists a lower triangular probability matrix \mathbf{P} that corresponds to a population satisfying the monotonicity assumption, i.e., $Y_i(1) \geq Y_i(0)$ for all i . Strassen (1965) and Rosenbaum (2001) demonstrated this result, and Theorem 1 extends the previous result without imposing the stochastic dominance assumption. Moreover, Theorem 1 also justifies the use of $\min_{0 \leq j \leq J-1} \Delta_j$ as a measure of the deviation from the stochastic dominance assumption (Scharfstein et al. 2004).

Next we consider bounding η . Realizing that $\eta = 1 - \text{pr}\{Y_i(0) \geq Y_i(1)\}$, we can derive bounds for $\text{pr}\{Y_i(0) \geq Y_i(1)\}$ by switching the treatment and control labels and applying Theorem 1.

Theorem 2. The sharp lower and upper bounds of η are

$$\eta_L = \max_{0 \leq j \leq J-1} \Delta_j, \quad \eta_U = 1 + \min_{0 \leq j \leq J-1} (\Delta_j - p_{j+}). \quad (5)$$

Deriving the sharp bounds of τ and η is related to a classical probability problem, first posed by A. N. Kolmogorov (c.f. Nelsen 2006): how to bound the distribution of the sum (or difference)

of two random variables with fixed marginal distributions? When $\delta = Y(1) - Y(0)$ is well-defined as for continuous outcomes, our causal parameters τ and η are determined by the distribution of the causal effect δ , the difference between the treatment and control potential outcomes. The sharp bounds on the distribution of δ have been obtained by Makarov (1982), Rüschendorf (1982) and Frank et al. (1987), and recently reviewed by Fan and Park (2010) and Fan et al. (2014). However, their results and proofs apply to the cases when $Y(1) - Y(0)$ is well defined, and the bounds in Theorems 1 and 2 hold in general including both continuous and ordinal outcomes.

In the proofs of Theorem 1 and 2 we construct the probability matrices that achieve the lower and upper bounds of τ and η , which correspond to negatively associated and positively associated potential outcomes. They are both extreme scenarios. In practice, researchers may also be interested in the case with independent potential outcomes (Rubin 1978; Cheng 2009; Agresti 2010; Ding and Dasgupta 2016), i.e., $p_{kl} = p_{k+}p_{+l}$ for all k and l . With independent potential outcomes, we can identify τ and η from the marginal distributions of the potential outcomes.

Theorem 3. With independent potential outcomes,

$$\tau_I = \sum_{k \geq l} \sum p_{k+p+l}, \quad \eta_I = \sum_{k > l} \sum p_{k+p+l}.$$

Furthermore, $\tau_L \leq \tau_I \leq \tau_U$ and $\eta_L \leq \eta_I \leq \eta_U$.

In cases where negatively associated potential outcomes are unlikely, we can use τ_I and η_I as the lower bounds of τ and η . Below we give two numerical examples to illustrate Theorems 1–3.

Example 3. The marginal probabilities $\mathbf{p}_1 = (1/5, 3/5, 1/5)^T$ and $\mathbf{p}_0 = (2/5, 1/5, 2/5)^T$ do not satisfy the stochastic dominance assumption, because $\Delta_0 = 0$, $\Delta_1 = 1/5 > 0$ and $\Delta_2 = -1/5 < 0$. Theorems 1 and 3 imply that $\tau_L = 2/5$, $\tau_I = 16/25$, and $\tau_U = 4/5$. The probability matrices corresponding to negatively associated, independent, and positively associated potential outcomes achieving these values are respectively

$$\mathbf{P}_1 = \begin{pmatrix} 0 & 1/5 & 0 \\ 1/5 & 0 & 2/5 \\ 2/5 & 0 & 0 \end{pmatrix}, \quad \mathbf{P}_2 = \begin{pmatrix} 2/25 & 1/25 & 2/25 \\ 6/25 & 3/25 & 6/25 \\ 2/25 & 1/25 & 2/25 \end{pmatrix}, \quad \mathbf{P}_3 = \begin{pmatrix} 1/5 & 0 & 0 \\ 1/5 & 1/5 & 1/5 \\ 0 & 0 & 1/5 \end{pmatrix}. \quad (6)$$

Similarly, Theorems 2 and 3 imply $\eta_L = 1/5$, $\eta_I = 9/25$, and $\eta_U = 3/5$.

Example 4. The marginal probabilities $\mathbf{p}_1 = (1/5, 1/5, 3/5)^T$ and $\mathbf{p}_0 = (3/5, 1/5, 1/5)^T$ satisfy the stochastic dominance assumption, because $\Delta_0 = 0$, $\Delta_1 = 2/5 > 0$ and $\Delta_2 = 2/5 > 0$. Theorems 1 and 3 imply $\tau_L = 3/5$, $\tau_I = 22/25$, and $\tau_U = 1$. The probability matrices corresponding to negatively associated, independent, and positively associated potential outcomes achieving these values are respectively

$$\mathbf{P}_4 = \begin{pmatrix} 0 & 1/5 & 0 \\ 0 & 0 & 1/5 \\ 3/5 & 0 & 0 \end{pmatrix}, \quad \mathbf{P}_5 = \begin{pmatrix} 3/25 & 1/25 & 1/25 \\ 3/25 & 1/25 & 1/25 \\ 9/25 & 3/25 & 3/25 \end{pmatrix}, \quad \mathbf{P}_6 = \begin{pmatrix} 1/5 & 0 & 0 \\ 0 & 1/5 & 0 \\ 2/5 & 0 & 1/5 \end{pmatrix}. \quad (7)$$

Similarly, Theorems 2 and 3 imply $\eta_L = 2/5$, $\eta_I = 3/5$, and $\eta_U = 4/5$.

As demonstrated in Examples 3 and 4, the bounds of τ (or η) generally do not shrink to a point. However, there are some special cases in which the lower and upper bounds of τ (or η) are identical. The sufficient and necessary conditions appear to be technical, and therefore we relegate the discussion to the Supplementary Material.

3.2. Covariate Adjustment

With pretreatment covariates, it is possible to further sharpen the bounds of the causal parameters (Grilli and Mealli 2008; Lee 2009; Long and Hudgens 2013; Mealli and Pacini 2013). Without loss of generality, we focus only on the bounds of τ . Within each level of the pretreatment covariates $\mathbf{X} = \mathbf{x}$,

$$\tau(\mathbf{x}) = \text{pr}\{Y(1) \geq Y(0) \mid \mathbf{X} = \mathbf{x}\}$$

is the conditional probability that the treatment is beneficial. We can obtain the conditional lower and upper bounds $\tau_L(\mathbf{x})$ and $\tau_U(\mathbf{x})$ given the covariate level \mathbf{x} , then average them over the covariate distribution $F(\mathbf{x})$, and finally obtain the adjusted bounds for τ :

$$\tau'_L = \int \tau_L(\mathbf{x}) F(d\mathbf{x}), \quad \tau'_U = \int \tau_U(\mathbf{x}) F(d\mathbf{x}). \quad (8)$$

Theorem 4. The adjusted bounds are tighter, i.e., $\tau_L \leq \tau'_L \leq \tau'_U \leq \tau_U$.

Theorem 4 holds intuitively, because the existence of covariates imposes more distributional restrictions on the observed data. We use the following example to illustrate Theorem 4.

Example 5. Consider a population consisting of two sub-populations of equal sizes, labeled by a binary covariate \mathbf{X} . Assume that the potential outcomes of sub-populations $\mathbf{X} = 1$ and $\mathbf{X} = 0$ are the independent potential outcomes in Example 3 and 4. Simple algebra gives the following joint distribution, marginal distributions, and τ of the potential outcomes:

$$\mathbf{P} = \begin{pmatrix} 1/10 & 1/25 & 3/50 \\ 9/50 & 2/25 & 7/50 \\ 11/50 & 2/25 & 1/10 \end{pmatrix}, \quad \mathbf{p}_1 = (1/5, 2/5, 2/5)^T, \quad \mathbf{p}_0 = (1/2, 1/5, 3/10)^T, \quad \tau = 19/25.$$

Without covariate information, Theorem 1 implies $\tau_L = 1/2$ and $\tau_U = 1$. However, if we first obtain the bounds for the two sub-populations and then average over them, we obtain sharper covariate adjusted bounds $\tau'_L = \tau_L(1)/2 + \tau_L(0)/2 = 1/2$, and $\tau'_U = \tau_U(1)/2 + \tau_U(0)/2 = 9/10$.

3.3. Identifying the Bounds from Observed Data

Previous subsections discussed the causal parameters τ and η and their bounds. The causal parameters depend on the joint distribution of the potential outcomes, but the bounds depend only on the marginal distributions of the potential outcomes. In practice, the observed data provide full information about only the marginal distributions. Therefore, point estimations of the bounds can be obtained, although the causal parameters themselves are only partially identified (c.f. Romano and Shaikh 2008, 2010; Richardson et al. 2014).

For unit $i = 1, \dots, N$, let the treatment indicator be Z_i , and the observed outcome be $Y_i^{\text{obs}} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. To avoid conceptual complications, we consider treatment assignments that satisfy the ignorability assumption (Rosenbaum and Rubin 1983), i.e., $Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid \mathbf{X}$. The ignorability assumption holds by the design of randomized experiments, and cannot be validated in observational studies. Under the ignorability assumption, we define the propensity score as $e(\mathbf{X}) = \text{pr}(Z = 1 \mid \mathbf{X})$, which is a constant independent of \mathbf{X} in completely randomized experiments. We

can identify the marginal distributions of the potential outcomes by

$$\text{pr}\{Y(1) = k\} = \text{E} \left\{ \frac{Z1(Y^{\text{obs}} = k)}{e(\mathbf{X})} \right\}, \quad \text{pr}\{Y(0) = l\} = \text{E} \left\{ \frac{(1 - Z)1(Y^{\text{obs}} = l)}{1 - e(\mathbf{X})} \right\}.$$

By replacing the expectations by their sample analogues, we obtain the moment estimators for the marginal distributions. We defer more detailed discussion about statistical inference to Section 5.

4. Randomized Experiments with Noncompliance

4.1. Causal Effects for Compliers

Noncompliance is an important topic in practice. For instance, in clinical trials some patients may not comply with their assigned treatments. Although noncompliance itself has been extensively investigated in the causal inference literature (e.g., Angrist et al. 1996), there appears to be very limited discussions about causal inference of ordinal outcomes in the presence of noncompliance. To the best of our knowledge, Cheng (2009) discussed various causal parameters under the assumptions of one-sided noncompliance, and Baker (2011) generalized her results to two-sided noncompliance; both of them assumed independent potential outcomes.

Under the Stable Unit Treatment Value Assumption, for unit i , let $\{D_i(1), D_i(0)\}$ be the potential values of treatment received under treatment and control; the observed treatment received is therefore $D_i^{\text{obs}} = Z_i D_i(1) + (1 - Z_i) D_i(0)$. Angrist et al. (1996) proposed to classify the units into four categories according to the joint values of $D_i(1)$ and $D_i(0)$:

$$G_i = \begin{cases} a, & \text{if } D_i(1) = 1, D_i(0) = 1, \\ c, & \text{if } D_i(1) = 1, D_i(0) = 0, \\ d, & \text{if } D_i(1) = 0, D_i(0) = 1, \\ n, & \text{if } D_i(1) = 0, D_i(0) = 0, \end{cases} \quad (9)$$

and referred to the subgroups defined in (9) as always-takers (a), compliers (c), defiers (d) and never-takers (n). Let $\pi_g = \text{pr}(G = g)$ denote the probability of the stratum $g \in \{a, c, d, n\}$, and

$$g_{kl} = \text{pr}\{Y(1) = k, Y(0) = l \mid G = g\}$$

be the probability of potential outcome k under treatment and potential outcome l under control within stratum g . The $J \times J$ probability matrix $\{g_{kl}\}_{0 \leq k, l \leq J-1}$ summarizes the joint distribution of the potential outcomes for stratum g . Define

$$g_{k+} = \sum_{l'=0}^{J-1} g_{kl'}, \quad g_{+l} = \sum_{k'=0}^{J-1} g_{k'l} \quad (k, l = 0, 1, \dots, J-1); \quad (10)$$

the vectors $(g_{0+}, \dots, g_{J-1,+})^T$ and $(g_{+0}, \dots, g_{+,J-1})^T$ characterize the marginal distributions of the potential outcomes under treatment and control. By the law of total probability,

$$p_{kl} = \sum_g \pi_g g_{kl}, \quad p_{k+} = \sum_g \pi_g g_{k+}, \quad p_{+l} = \sum_g \pi_g g_{+l}. \quad (11)$$

We define the subgroup causal parameters within stratum g as

$$\tau_g = \text{pr} \{Y_i(1) \geq Y_i(0) \mid G = g\} = \sum_{k \geq l} g_{kl}, \quad \eta_g = \text{pr} \{Y_i(1) > Y_i(0) \mid G = g\} = \sum_{k > l} g_{kl}.$$

Following Angrist et al. (1996), we invoke the following assumptions: (1) Complete Randomization, i.e., $Z \perp\!\!\!\perp \{D(1), D(0), Y(1), Y(0), \mathbf{X}\}$; (2) Strong Monotonicity, i.e., $D_i(0) = 0$ for all i , or Monotonicity, i.e., $D_i(1) \geq D_i(0)$ for all i ; (3) Exclusion Restriction, i.e., $D_i(1) = D_i(0)$ implies $Y_i(1) = Y_i(0)$. Monotonicity rules out the defiers with $G = d$, and strong monotonicity further rules out the always-takers with $G = a$. Exclusion restriction implies that $\tau_n = 1, \eta_n = 0, \tau_a = 1$ and $\eta_a = 0$. Therefore, we discuss only the causal effects for the compliers, i.e., τ_c and η_c .

4.2. Bounds on the Causal Effects for Compliers

We focus only on the case under monotonicity, because it is more general than strong monotonicity. Under monotonicity and exclusion restriction, we can identify the probabilities of always-takers, compliers and never-takers, i.e., (π_a, π_c, π_n) , and the distributions of the potential outcomes conditional on G (Angrist et al. 1996; Cheng 2009; Baker 2011), i.e., the g_{k+} 's and g_{+l} 's. Below, we establish the relationships between the causal parameters τ and τ_c , and between η and η_c .

Theorem 5. $\tau_c = \tau / \pi_c - (1 - \pi_c) / \pi_c$ and $\eta_c = \eta / \pi_c$.

Therefore, we can plug in the upper and lower bounds of τ and η to obtain the bounds of τ_c and

η_c , using the relationships in Theorem 5. However, these bounds are not sharp, and the following bounds, implied by Theorems 1 and 2, are narrower.

Corollary 2. The sharp lower and upper bounds of τ_c are

$$\tau_{c,L} = \max_{0 \leq j \leq J-1} (c_{+j} + \Delta_{c,j}), \quad \tau_{c,U} = 1 + \min_{0 \leq j \leq J-1} \Delta_{c,j},$$

and the sharp lower and upper bounds of η_c are

$$\eta_{c,L} = \max_{0 \leq j \leq J-1} \Delta_{c,j}, \quad \eta_{c,U} = 1 + \min_{0 \leq j \leq J-1} (\Delta_{c,j} - c_{j+}).$$

Similar to Section 3.2, we can use covariates to sharpen the bounds of τ_c . Within each level of the pretreatment covariates $\mathbf{X} = \mathbf{x}$, we define the conditional probabilities that the treatment is beneficial for compliers as

$$\tau_c(\mathbf{x}) = \text{pr}\{Y(1) \geq Y(0) \mid G = c, \mathbf{X} = \mathbf{x}\},$$

and obtain their conditional sharp upper and lower bounds $\tau_{c,L}(\mathbf{x})$ and $\tau_{c,U}(\mathbf{x})$. Because

$$\tau_c = \frac{\int \tau_c(\mathbf{x}) \pi_c(\mathbf{x}) dF(\mathbf{x})}{\int \pi_c(\mathbf{x}) dF(\mathbf{x})},$$

the bounds for τ_c become

$$\tau'_{c,L} = \frac{\int \tau_{c,L}(\mathbf{x}) \pi_c(\mathbf{x}) dF(\mathbf{x})}{\int \pi_c(\mathbf{x}) dF(\mathbf{x})}, \quad \tau'_{c,U} = \frac{\int \tau_{c,U}(\mathbf{x}) \pi_c(\mathbf{x}) dF(\mathbf{x})}{\int \pi_c(\mathbf{x}) dF(\mathbf{x})}. \quad (12)$$

Similar to Theorem 4, the adjusted bounds are tighter, i.e., $\tau_{c,L} \leq \tau'_{c,L} \leq \tau'_{c,U} \leq \tau_{c,U}$.

4.3. Using Noncompliance to Sharpen Bounds for the Whole Population

Theorem 5 and Corollary 2 imply two new sets of bounds for τ and η , which are tighter than those in Theorems 1 and 2.

Corollary 3. We can bound τ from below and above using

$$\tau_L'' = \pi_c \tau_{c,L} + 1 - \pi_c, \quad \tau_U'' = \pi_c \tau_{c,U} + 1 - \pi_c,$$

and bound η from below and above using

$$\eta_L'' = \pi_c \eta_{c,L}, \quad \eta_U'' = \pi_c \eta_{c,U}.$$

These new bounds above are narrower than those in Theorems 1 and 2, because they satisfy $\tau_L \leq \tau_L'', \tau_U = \tau_U'', \eta_L = \eta_L'',$ and $\eta_U \geq \eta_U''$.

There are two reasons that we can obtain tighter bounds. First, we use the partially observed variable G as a pretreatment variable. Second, the monotonicity and exclusion restriction assumptions further restrict the probability structure of the potential outcomes.

5. Statistical Inference of the Bounds

In practice, we need to estimate the marginal probabilities of the potential outcomes and the bounds. To save space for the main text, we discuss only the bounds of τ and τ_c .

5.1. Point Estimation

Completely Randomized Experiments To estimate the unadjusted bounds, we replace p_{k+} and p_{+l} in Theorem 1 with their sample analogues. Moreover, to estimate the covariate adjusted bounds in (8), we invoke parametric models such as proportional odds models to estimate the marginal probabilities of the potential outcomes of unit i as $\hat{p}_{k+}(\mathbf{x}_i)$ and $\hat{p}_{+l}(\mathbf{x}_i)$, and use them to estimate the sharp lower and upper bounds, $\hat{\tau}_L(\mathbf{x}_i)$ and $\hat{\tau}_U(\mathbf{x}_i)$, for $\tau(\mathbf{x}_i)$. Finally, the estimated adjusted bounds of τ are

$$\hat{\tau}_L' = N^{-1} \sum_{i=1}^N \hat{\tau}_L(\mathbf{x}_i), \quad \hat{\tau}_U' = N^{-1} \sum_{i=1}^N \hat{\tau}_U(\mathbf{x}_i).$$

Unconfounded Observational Studies If we have propensity score estimator $\widehat{e}(\mathbf{x}_i)$ for unit i , then we can estimate the marginal probabilities by

$$\widehat{p}_{k+} = N^{-1} \sum_{i=1}^N Z_i \frac{1(Y_i^{\text{obs}} = k)}{\widehat{e}(X_i)}, \quad \widehat{p}_{+l} = N^{-1} \sum_{i=1}^N (1 - Z_i) \frac{1(Y_i^{\text{obs}} = l)}{1 - \widehat{e}(X_i)},$$

and then estimate the bounds accordingly.

Completely Randomized Experiments With Noncompliance Without covariates, we use the EM algorithm (Dempster et al. 1977) to estimate π_c , c_{k+} and c_{+l} , and then estimate the unadjusted bounds in Corollary 2. For a more detailed description of the EM algorithm, see Baker (2011). With covariates, we need to invoke parametric models for G (e.g., multinomial logistic model given \mathbf{X}) and the marginal probabilities of the potential outcomes, and use the EM algorithm to compute the maximum likelihood of the model parameters (cf. Zhang et al. 2009; Frumento et al. 2012). After obtaining the sample analogues of $\tau_{c,L}(\mathbf{x})$, $\tau_{c,U}(\mathbf{x})$ and $\pi_c(\mathbf{x})$, we can estimate the covariate adjusted bounds defined in (12) using a plug-in approach.

5.2. Confidence Intervals

To quantify the uncertainty associated with the aforementioned estimators of the bounds, we can use the bootstrap method proposed by Horowitz and Manski (2000) to obtain the confidence intervals (CI) for the unadjusted and covariate adjusted bounds. For computational details of some other bootstrap methods, see Cheng and Small (2006) and Yang and Small (2016).

Because the upper and lower bounds involve maximum and minimum of several terms, their asymptotic distributions are not normal, and the construction of confidence intervals on the bounds becomes challenging (Hirano and Porter 2012). Recently, Romano and Shaikh (2008, 2010) and Chernozhukov et al. (2013) proposed some delicate methods to construct confidence intervals for partially identified parameters. However, several researchers (e.g., Cheng and Small 2006; Fan and Park 2010; Yang 2014) evaluated the performance of the bootstrapped confidence intervals for partially identified parameters (Beran 1988, 1990; Horowitz and Manski 2000) via extensive simulations. Although the rigorous theoretical guarantee of the bootstrapped confidence intervals has not been fully established, they found the bootstrapped confidence intervals work fairly well in various settings,

and their performances are at least comparable to the more delicate methods mentioned before. Therefore, for simplicity in simulations and transparency in applications, we still use bootstrap to construct confidence intervals. We provide the code for implementation, and more sophisticated users can modify our code to include the more advanced methods.

6. Simulation Studies

6.1. Without Noncompliance

To save space in the main text, we focus only on τ and its bounds in Theorem 1. We choose the sample size to be 200, and consider four cases with different probability matrices \mathbf{P} 's. Cases 1 and 2 correspond to matrices \mathbf{P}_2 and \mathbf{P}_3 in (6), i.e., the independent and positively associated potential outcomes, which share the same marginal distribution but do not satisfy the stochastic dominance assumption. Cases 3 and 4 correspond to matrices \mathbf{P}_5 and \mathbf{P}_6 in (7), i.e., the independent and positively associated potential outcomes, which share the same marginal distribution and satisfy the stochastic dominance assumption. Columns 2–4 of Table 1 summarize the true values of τ , τ_L and τ_U , for all four cases. For Cases 1 and 3 with independent potential outcomes, $\tau_L < \tau < \tau_U$. For Cases 2 and 4 with positively associated potential outcomes, $\tau = \tau_U$.

For each case, we independently draw 5000 treatment assignments from a balanced completely randomized experiment. For each observed dataset, we calculate point estimates of τ_L and τ_U , and construct a 95% confidence interval for the bounds (τ_L, τ_U) , i.e., a confidence interval that contains both the lower bound τ_L and the upper bound τ_U with probability 0.95. In columns 5–8 of Table 1, we report the biases and standard errors of the point estimators $\hat{\tau}_L$ and $\hat{\tau}_U$; in columns 9 and 10 of Table 1, we report the coverage rates of the intervals on the bounds (τ_L, τ_U) and the true parameter τ . Table 1 shows that the point estimators have small biases and standard errors, and the confidence intervals achieve reasonable coverage rates on the bounds (τ_L, τ_U) , although they over-cover the true parameter τ .

6.2. With Noncompliance

To evaluate the finite sample performances of the estimators and the confidence intervals of the bounds, we conduct simulation studies under different model specifications. To save space, we

Table 1: Numerical examples without noncompliance. The first three columns contain the true values, the next four columns contain the biases and standard errors of the point estimators of the bounds, and the last two columns contain the coverage properties of the confidence intervals for the bounds and the true parameter.

Case	τ	τ_L	τ_U	bias $_L$	se $_L$	bias $_U$	se $_U$	coverage $_1$	coverage $_2$
1	0.640	0.400	0.800	0.016	0.037	0.000	0.045	0.987	1.000
2	0.800	0.400	0.800	0.013	0.043	-0.001	0.057	0.957	0.974
3	0.880	0.600	1.000	0.026	0.030	0.000	0.000	0.967	1.000
4	1.000	0.600	1.000	0.025	0.031	0.000	0.000	0.960	1.000

focus only on the parameter τ_c , and consider six simulation cases. Cases 1–3 are indexed by the parameter $\beta \in \{1, 1/2, 0\}$, and Cases 4–6 are indexed by the parameter $\xi \in \{1, 1/2, 0\}$. We postpone the interpretations of β and η until afterwards. For each case, let the pretreatment covariates $\mathbf{X} = (1, X_1, X_2)$, where $X_1 \sim N(0, 1)$, and $X_2 \sim \text{Bern}(1/2)$. For fixed $\mathbf{X} = \mathbf{x}$, we generate the variable G from a multiple logistic model:

$$\pi_g(\mathbf{x}) = \exp(\boldsymbol{\eta}_g^T \mathbf{x}) / \left\{ \sum_{g'} \exp(\boldsymbol{\eta}_{g'}^T \mathbf{x}) \right\} \quad (g = a, c, n),$$

where $\boldsymbol{\eta}_c = \mathbf{0}$, $\boldsymbol{\eta}_a = (1/2, 1, 0)$ and $\boldsymbol{\eta}_n = (-1/2, 1, 0)$. We generate the potential outcomes from proportional odds models.

1. For always-takers, let $Y_i(1) = Y_i(0)$, and their marginal distributions be

$$\text{logit} \left\{ \sum_{k \leq j} a_{k+}(\mathbf{x}) \right\} = \text{logit} \left\{ \sum_{l \leq j} a_{+l}(\mathbf{x}) \right\} = \alpha_{a,j} - 2x_1,$$

where $\alpha_{a,0} = -1/2$ and $\alpha_{a,1} = 1$.

2. For never-takers let $Y_i(1) = Y_i(0)$, and their marginal distributions be

$$\text{logit} \left\{ \sum_{k \leq j} n_{k+}(\mathbf{x}) \right\} = \text{logit} \left\{ \sum_{l \leq j} n_{+l}(\mathbf{x}) \right\} = \alpha_{n,j},$$

where $\alpha_{a,0} = -3/2$ and $\alpha_{a,1} = 0$.

3. For compliers let $Y_i(1)$ and $Y_i(0)$ be independent, and the values of the parameters be $\alpha_{c,0} = -1$, $\alpha_{c,1} = 1/2$, $\gamma_{c,0} = 1/2$ and $\gamma_{c,1} = 2$.

(a) For Cases 1–3, let the marginal distributions be

$$\text{logit} \left\{ \sum_{k \leq j} c_{k+}(\mathbf{x}) \right\} = \alpha_{c,j} - 2\beta x_1, \quad \text{logit} \left\{ \sum_{l \leq j} c_{+l}(\mathbf{x}) \right\} = \gamma_{c,j} + \beta x_1;$$

(b) For Cases 4–6, let the marginals distributions be

$$\text{logit} \left\{ \sum_{k \leq j} c_{k+}(\mathbf{x}) \right\} = \alpha_{c,j} - 2x_1 - \xi x_2, \quad \text{logit} \left\{ \sum_{l \leq j} c_{+l}(\mathbf{x}) \right\} = \gamma_{c,j} + x_1 + \xi x_2.$$

For the above six cases, their true values of τ_c , unadjusted and adjusted bounds are in columns 2–4 of each sub-table of Table 2. For Cases 1–3, the parameter β quantifies the association between the covariates and the potential outcomes. As β decreases, the covariate adjusted bounds become closer to the unadjusted bounds. For Cases 4–6, the parameter ξ quantifies the association between the binary covariate X_2 and the potential outcomes of compliers.

We conduct inference without the binary covariate X_2 . This does not affect Cases 1–3 because X_2 is irrelevant in the data generating process, but does affect Cases 4–6. We purposefully design the data generating process in this way, to examine the performance of our estimators under correct and incorrect model specifications. For each case, we choose the sample size to be 1000, and independently draw 1000 treatment assignments from a balanced completely randomized experiment. For each observed dataset, we first estimate the bounds $\tau_{c,L}$ and $\tau_{c,U}$, and construct a 95% confidence interval for $(\tau_{c,L}, \tau_{c,U})$; we then estimate the bounds $\tau'_{c,L}$ and $\tau'_{c,U}$, and construct a 95% confidence interval for $(\tau'_{c,L}, \tau'_{c,U})$.

We report the simulation results in Table 2, in which columns 4–7 of each sub-table include the biases of the point estimators, the average lengths and coverage rates of the 95% confidence intervals on the bounds. First, the point estimators of the bounds have small biases. Second, when the pretreatment covariates are associated with the potential outcomes, the confidence intervals of the bounds $(\tau_{c,L}, \tau_{c,U})$ are longer than those of $(\tau'_{c,L}, \tau'_{c,U})$, on average. Third, the confidence intervals for the bounds $(\tau_{c,L}, \tau_{c,U})$ and $(\tau'_{c,L}, \tau'_{c,U})$ achieve reasonable coverage rates. Fourth, the performance of the bounds is robust to the missing of the binary covariate, or, equivalently, misspecification of the outcome models.

Table 2: Numerical examples with noncompliance. In each sub-table, the first three columns contain the true values of the causal parameter τ_c and its lower and upper bounds, the next two columns contain the biases of the point estimators of the lower and upper bounds, and the last two columns contain the lengths and coverage rates of the 95% confidence intervals for the bounds.

(a) Unadjusted Bounds

Case	τ_c	$\tau_{c,L}$	$\tau_{c,U}$	bias _L	bias _U	length	coverage
1	0.686	0.488	0.970	0.002	-0.028	0.658	0.947
2	0.770	0.553	1.000	0.005	-0.005	0.574	0.973
3	0.856	0.622	1.000	0.034	-0.000	0.485	0.958
4	0.782	0.590	1.000	0.000	-0.002	0.528	0.976
5	0.738	0.542	1.000	0.002	-0.016	0.588	0.966
6	0.686	0.488	0.970	0.002	-0.028	0.658	0.947

(b) Adjusted Bounds

Case	τ_c	$\tau'_{c,L}$	$\tau'_{c,U}$	bias _L	bias _U	length	coverage
1	0.686	0.503	0.772	0.008	-0.006	0.466	0.968
2	0.770	0.563	0.935	0.004	-0.007	0.530	0.968
3	0.856	0.622	1.000	0.021	-0.001	0.489	0.959
4	0.782	0.602	0.846	0.005	0.012	0.436	0.960
5	0.738	0.556	0.817	0.007	-0.003	0.447	0.965
6	0.686	0.503	0.772	0.008	-0.006	0.466	0.968

7. Applications

7.1. A Taste-Testing Experiment Without Noncompliance

We use the taste-testing experiment data in Bradley et al. (1962) to demonstrate the estimation and inference of the proposed causal parameters. The outcome of interest Y is ordinal with five categories, from “terrible” with $Y = 0$ to “excellent” with $Y = 4$. We consider only three treatments C, D, E, and summarize the data and results in Table 3. The negative associated potential outcomes appear unlikely for this example, therefore we focus on the interpretations of the cases with independent and positive correlated potential outcomes, e.g., τ_I and τ_U . First, treatment E stochastically dominates treatment C, and the confidence intervals for (τ_I, τ_U) and (η_I, η_U) are (0.913, 1.000) and (0.651, 1.000). The results suggest that treatment E is indeed better than treatment C, because both lower confidence limits are greater than 0.5. Second, although treatment E and treatment D do not stochastically dominate each other, the confidence intervals for (τ_I, τ_U) and (η_I, η_U) are (0.656, 0.982) and (0.519, 0.886), suggesting that treatment E is better than treatment D. Therefore the proposed causal parameters τ and η are useful for decision making, especially when the stochastic dominance assumption does not hold.

Table 3: Analysis of a Taste-Testing Experiment

(a) Data from Bradley et al. (1962)

Treatment	Outcome Categories					row sum
	0	1	2	3	4	
C	14	13	6	7	0	40
D	11	15	3	5	8	42
E	0	2	10	30	2	44

(b) Results for τ

	$\hat{\tau}_L$	$\hat{\tau}_I$	$\hat{\tau}_U$	CI for (τ_L, τ_U)	CI for (τ_I, τ_U)
E vs C	0.779	0.945	1.000	(0.673, 1.000)	(0.913, 1.000)
E vs D	0.645	0.782	0.855	(0.495, 1.000)	(0.656, 0.982)

(c) Results for η

	$\hat{\eta}_L$	$\hat{\eta}_I$	$\hat{\eta}_U$	CI for (η_L, η_U)	CI for (η_I, η_U)
E vs C	0.630	0.777	0.870	(0.480, 1.000)	(0.651, 1.000)
E vs D	0.574	0.660	0.736	(0.423, 0.886)	(0.519, 0.886)

7.2. A Job Training Program with Noncompliance

In the mid-1990s, Mathematica Policy Research conducted an experiment that randomly enrolled eligible applicants into the Job Corps program (Schochet et al. 2003; Lee 2009). We re-analyzed the dataset from 1995 with 13499 units. For detailed descriptions of the dataset, see Zhang et al. (2009) and Frumento et al. (2012). In our analysis, $Z = 1$ if an applicant was enrolled in the program, and $Z = 0$ otherwise; $D = 1$ if an applicant actually participated in the program, and $D = 0$ otherwise. The strong monotonicity assumption holds by design. Using the hourly wage after 52 weeks of enrollment, we create a three-level ordinal outcome Y as follows: $Y = 0$ for zero wage because of unemployment, $Y = 1$ for low wage (no more than 4.25 U.S dollars, 150 % of the minimal wage at the time the data was collected), and $Y = 2$ for high wage (more than 4.25 U.S dollars). The covariates include gender, age, education, marital status, etc.

Table 4 summarizes the results. For both causal parameters τ_c and η_c , the confidence intervals for the lower and upper bounds become narrower when we take covariates into account. Similarly as the previous example, we focus on the interpretations of the cases with independent and positive correlated potential outcomes. The confidence intervals with or without covariates for (τ_I, τ_U) suggest that the hourly wages of more than 70% of participants does not decrease because of the job training program. Additionally, the confidence intervals with or without covariates for (η_I, η_U) suggest that the hourly wages of roughly 20%–30% of participants strictly increase because of the job training program.

Table 4: Analysis of the Job Corps Program

(a) Results for τ					
	$\hat{\tau}_{c,L}$	$\hat{\tau}_{c,I}$	$\hat{\tau}_{c,U}$	CI for $(\tau_{c,L}, \tau_{c,U})$	CI for $(\tau_{c,I}, \tau_{c,U})$
w/o Covariates	0.561	0.708	0.913	(0.538, 0.937)	(0.687, 0.934)
w/ Covariates	0.592	0.722	0.910	(0.571, 0.931)	(0.701, 0.931)

(b) Results for η					
	$\hat{\eta}_{c,L}$	$\hat{\eta}_{c,I}$	$\hat{\eta}_{c,U}$	CI for $(\eta_{c,L}, \eta_{c,U})$	CI for $(\eta_{c,I}, \eta_{c,U})$
w/o Covariates	0.005	0.209	0.352	(0.000, 0.361)	(0.199, 0.362)
w/ Covariates	0.006	0.193	0.319	(0.000, 0.329)	(0.181, 0.330)

As a final note, we use this example to illustrate Corollary 3. Without the noncompliance information, the estimators of the bounds of τ are $\hat{\tau}_L = 0.558$ and $\hat{\tau}_U = 0.937$, with 95% confidence

interval (0.542, 0.953); the estimators of the bounds of η are $\hat{\eta}_L = 0.004$ and $\hat{\eta}_U = 0.379$, with 95% confidence interval (0.000, 0.388). With the noncompliance information, the estimators of the bounds of τ are $\hat{\tau}_L'' = 0.683$ and $\hat{\tau}_U'' = 0.937$, with 95% confidence interval (0.667, 0.953); the estimator of the bounds of η are $\hat{\eta}_L'' = 0.004$ and $\hat{\eta}_U'' = 0.254$, with 95% confidence interval (0.000, 0.262). Therefore, the noncompliance information in return improves the inference of τ and η for the whole population.

8. Concluding Remarks

We proposed to use two causal parameters to evaluate treatment effect on ordinal outcomes, and derived the explicit forms of their sharp bounds by using only the marginal distributions of the potential outcomes. Although we advocate the use of parameters τ and η to measure treatment effects, we acknowledge that some other causal parameters may also provide some information in practice (e.g. Agresti 2010; Volfovsky et al. 2015). For general parameters, although deriving the explicit forms of the bounds may be difficult, we may use numerical methods. For instance, we can use numerical linear programs to calculate the maximum and minimum values of the relative treatment effect $\alpha = \tau + \eta - 1$ under the constraints in (4).

Appendix

We first state a lemma extending a result in Strassen (1965). This lemma plays a central role in our later proofs, and is also of independent interest. We provide the proof of the lemma in the Supplementary Material. In this Appendix, we present only the proof of Theorem 1, and relegate the proofs of other theorems and corollaries to the Supplementary Material.

Lemma 1. Assume that (x_0, \dots, x_{n-1}) and (y_0, \dots, y_{n-1}) are nonnegative constants.

- (a) If $\sum_{r=s}^{n-1} x_r \geq \sum_{r=s}^{n-1} y_r$ for all $s = 0, \dots, n-1$, there exists an $n \times n$ lower triangular matrix

$\mathbf{A}_n = (a_{kl})_{0 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} a_{kl'} \leq x_k, \quad \sum_{k'=0}^{n-1} a_{k'l} = y_l \quad (k, l = 0, \dots, n-1). \quad (13)$$

- (b) If $\sum_{r=s}^{n-1} x_r \leq \sum_{r=s}^{n-1} y_r$ for all $s = 0, \dots, n-1$, there exists an $n \times n$ upper triangular matrix $\mathbf{B}_n = (b_{kl})_{0 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} b_{kl'} = x_k, \quad \sum_{k'=0}^{n-1} b_{k'l} \leq y_l \quad (k, l = 0, \dots, n-1). \quad (14)$$

- (c) If $\sum_{r=0}^s x_r \leq \sum_{r=0}^s y_r$ for all $s = 0, \dots, n-1$, there exists an $n \times n$ lower triangular matrix $\mathbf{C}_n = (p_{kl})_{0 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} p_{kl'} = x_k, \quad \sum_{k'=0}^{n-1} p_{k'l} \leq y_l \quad (k, l = 0, \dots, n-1). \quad (15)$$

- (d) If $\sum_{r=0}^s x_r \geq \sum_{r=0}^s y_r$ for all $s = 0, \dots, n-1$, there exists an $n \times n$ upper triangular matrix $\mathbf{D}_n = (d_{kl})_{0 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{n-1} d_{kl'} \leq x_k, \quad \sum_{k'=0}^{n-1} d_{k'l} = y_l \quad (k, l = 0, \dots, n-1). \quad (16)$$

- (e) If we further assume $\sum_{r=0}^{n-1} y_r = \sum_{r=0}^{n-1} x_r$, the above inequalities in (13)–(16) all reduce to equalities, i.e., the matrices \mathbf{A}_n , \mathbf{B}_n , \mathbf{C}_n and \mathbf{D}_n have (x_0, \dots, x_{n-1}) and (y_0, \dots, y_{n-1}) as their row and column sums.

Proof of Theorem 1. For all $j = 0, 1, \dots, J-1$,

$$\begin{aligned} \tau &= \sum_{k \geq l} p_{kl} = 1 - \sum_{k < l} p_{kl} \\ &\leq 1 - \sum_{k < j} \sum_{l \geq j} p_{kl} = 1 - \left(\sum_{k=0}^{j-1} \sum_{l \geq j} p_{kl} - \sum_{k \geq j} \sum_{l \geq j} p_{kl} \right) \end{aligned} \quad (17)$$

$$\begin{aligned} &\leq 1 - \left(\sum_{k=0}^{j-1} \sum_{l \geq j} p_{kl} - \sum_{k \geq j} \sum_{l=1}^{j-1} p_{kl} \right) = 1 - \left(\sum_{l \geq j} p_{+l} - \sum_{k \geq j} p_{k+} \right) \\ &= 1 + \Delta_j, \end{aligned} \quad (18)$$

and

$$\begin{aligned}
\tau &= \sum_{k \geq l} \sum p_{kl} \\
&\geq \sum_{k \geq j} \sum_{l \leq j} p_{kl} = \sum_{k \geq j} \sum_{l=0}^{J-1} p_{kl} - \sum_{k \geq j} \sum_{l > j} p_{kl} \tag{19}
\end{aligned}$$

$$\begin{aligned}
&\geq \sum_{k \geq j} \sum_{l=0}^{J-1} p_{kl} - \sum_{k=0}^{J-1} \sum_{l > j} p_{kl} = \sum_{k \geq j} p_{k+} - \sum_{l > j} p_{+l} \tag{20} \\
&= p_{+j} + \Delta_j,
\end{aligned}$$

which implies that $\tau_L \leq \tau \leq \tau_U$.

We now construct two probability matrices attaining the lower and upper bounds respectively, using Lemma 1.

We first construct a probability matrix attaining the upper bound τ_U . Let

$$j_1 = \min \left\{ 0 \leq j' \leq J-1 : \Delta_{j'} = \min_{0 \leq j \leq J-1} \Delta_j \right\}$$

be the minimum index j that attains the minimum value of Δ_j 's. To attain τ_U , the equalities in (17) and (18) must hold, i.e.,

$$\sum_{k < l} p_{kl} = \sum_{k < j_1} \sum_{l \geq j_1} p_{kl}, \quad \sum_{k \geq j_1} \sum_{l \geq j_1} p_{kl} = \sum_{k \geq j_1} \sum_{l=1}^{J-1} p_{kl}. \tag{21}$$

If $j_1 = 0$, $\min_{0 \leq j \leq J-1} \Delta_j = \Delta_0 = 0$, implying that $\Delta_j = \sum_{k=j}^{J-1} p_{k+} - \sum_{l=j}^{J-1} p_{+l} \geq 0$ for all j , i.e., the marginal probabilities satisfy the stochastic dominance assumption. According to Lemma 1(e), there exists a lower triangular probability matrix \mathbf{P} with marginal probabilities $\mathbf{p}_1 = (p_{0+}, \dots, p_{J-1,+})^T$ and $\mathbf{p}_0 = (p_{+0}, \dots, p_{+,J-1})^T$. Correspondingly, $\tau = 1 + \Delta_0 = 1$.

If $j_1 > 0$, the constraints in (21) force some elements of the probability matrix to be zeros. To be more specific, the constraints in (21) imply that the probability matrix has the following block structure:

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{\text{tl}} & \mathbf{P}_{\text{tr}} \\ \mathbf{0} & \mathbf{P}_{\text{br}} \end{pmatrix}, \tag{22}$$

where the $j_1 \times j_1$ sub-matrix \mathbf{P}_{tl} on top left and the $(J - j_1) \times (J - j_1)$ sub-matrix \mathbf{P}_{br} on bottom right are both lower triangular, and the $j_1 \times (J - j_1)$ sub-matrix \mathbf{P}_{tr} on top right has no restrictions.

Because $\Delta_{j_1} \leq \Delta_j$ for all $j = 0, 1, \dots, J - 1$, we have

$$\sum_{k=j}^{j_1-1} p_{k+} \geq \sum_{l=j}^{j_1-1} p_{+l} \quad (j = 0, \dots, j_1 - 1); \quad \sum_{k=j_1}^j p_{k+} \leq \sum_{l=j_1}^j p_{+l} \quad (j = j_1, \dots, J - 1).$$

Given the above two sets of constraints on the marginal probabilities, we construct the probability matrix \mathbf{P} in three steps.

- (1) We apply Lemma 1(a) to $(p_{0+}, \dots, p_{j_1-1,+})$ and $(p_{+0}, \dots, p_{+,j_1-1})$, and obtain a lower triangular matrix $\mathbf{P}_{\text{tl}} = (p_{kl})_{0 \leq k, l \leq j_1-1}$ with nonnegative elements such that

$$\sum_{l'=0}^{j_1-1} p_{kl'} \leq p_{k+}, \quad \sum_{k'=0}^{j_1-1} p_{k'l} = p_{+l} \quad (k, l = 0, \dots, j_1 - 1).$$

- (2) We apply Lemma 1(c) to $(p_{j_1+}, \dots, p_{J-1,+})$ and $(p_{+j_1}, \dots, p_{+,J-1})$, and obtain a lower triangular matrix $\mathbf{P}_{\text{br}} = (p_{kl})_{j_1 \leq k, l \leq J-1}$ with nonnegative elements such that

$$\sum_{l'=j_1}^{J-1} p_{kl'} = p_{k+}, \quad \sum_{k'=j_1}^{J-1} p_{k'l} \leq p_{+l} \quad (k, l = j_1, \dots, J - 1).$$

- (3) We construct $\mathbf{P}_{\text{tr}} = (p_{kl})_{0 \leq k \leq j_1-1, j_1 \leq l \leq J-1}$ by letting

$$p_{kl} = \left(p_{k+} - \sum_{l'=0}^{j_1-1} p_{kl'} \right) \left(p_{+l} - \sum_{k'=j_1}^{J-1} p_{k'l} \right) \geq 0 \quad (k = 0, \dots, j_1 - 1; l = j_1, \dots, J - 1).$$

The constructed probability matrix \mathbf{P} has marginal probabilities $\mathbf{p}_1 = (p_{0+}, \dots, p_{J-1,+})^T$ and $\mathbf{p}_0 = (p_{+0}, \dots, p_{+,J-1})^T$. What is more, by (22) the τ of \mathbf{P} is the sum of all the elements in \mathbf{P}_{tl} and \mathbf{P}_{br} , which we construct in the above (1) and (2). Therefore, we have

$$\tau = \sum_{l'=0}^{j_1-1} p_{+l'} + \sum_{k'=j_1}^{J-1} p_{k'+} = 1 + \Delta_{j_1},$$

which implies that the probability matrix \mathbf{P} attains τ_U .

We then construct a probability matrix attaining the lower bound in τ_L . Let

$$j_2 = \min \left\{ j' : p_{+j'} + \Delta_{j'} = \max_{0 \leq j \leq J-1} (p_{+j} + \Delta_j) \right\}$$

be the minimum index j that attains the maximum value of $(p_{+j} + \Delta_j)$'s. To attain τ_L , the equalities in (19) and (20) must hold, i.e.,

$$\sum_{k \geq l} \sum p_{kl} = \sum_{k \geq j_2} \sum_{l \leq j_2} p_{kl}, \quad \sum_{k \geq j_2} \sum_{l > j_2} p_{kl} = \sum_{k=0}^{J-1} \sum_{l > j_2} p_{kl}. \quad (23)$$

If $j_2 = 0$, from (23) we know that the elements in the lower triangular part but not in the first column of the probability matrix \mathbf{P} are all zeros, i.e.,

$$\mathbf{P} = \begin{pmatrix} \mathbf{p} & \mathbf{P}_{\text{tr}} \\ p_{J-1,0} & 0^T \end{pmatrix}, \quad (24)$$

where $\mathbf{p} = (p_{0,0}, \dots, p_{J-2,0})^T$, and the $(J-1) \times (J-1)$ sub-matrix \mathbf{P}_{tr} on top right is upper triangular. Because $p_{+0} + \Delta_0 \geq p_{+j} + \Delta_j$ for all j , we have

$$\sum_{k=0}^j p_{k+} \geq \sum_{l=0}^j p_{+,l+1} \quad (j = 0, \dots, J-2).$$

Applying Lemma 1(d) to $(p_{0+}, \dots, p_{J-2,+})$ and $(p_{+1}, \dots, p_{+,J-1})$, we obtain an upper triangular matrix $\mathbf{P}_{\text{tr}} = (p_{kl})_{0 \leq k \leq J-2, 1 \leq l \leq J-1}$ with nonnegative elements such that

$$\sum_{l'=1}^{J-1} p_{kl'} \leq p_{k+}, \quad \sum_{k'=0}^{J-2} p_{k'l} = p_{+l} \quad (k = 0, \dots, J-2; l = 1, \dots, J-1).$$

To complete the construction, let $p_{J-1,0} = p_{J-1,+}$, and

$$p_{k0} = p_{k+} - \sum_{l'=1}^{J-1} p_{kl'} \geq 0 \quad (k = 0, \dots, J-2).$$

The constructed probability matrix \mathbf{P} has marginal probabilities $\mathbf{p}_1 = (p_{0+}, \dots, p_{J-1,+})^T$ and $\mathbf{p}_0 = (p_{+0}, \dots, p_{+,J-1})^T$. Moreover, by (24) the τ of \mathbf{P} is the sum of all the elements in the first

column. Therefore $\tau = p_{+0} = p_{+0} + \Delta_0$, which implies that \mathbf{P} attains τ_L .

If $j_2 = J - 1$, the proof is similar to the above case with $j_2 = 0$. If $0 < j_2 < J - 1$, because the first equality in (23) is equivalent to

$$\sum_{k < j_2} \sum_{l \leq k} p_{kl} + \sum_{k \geq j_2} \sum_{l \leq k} p_{kl} = \sum_{k \geq j_2} \sum_{l \leq j_2} p_{kl},$$

the probability matrix \mathbf{P} must satisfy the following constraints:

(C1) For all $k = 0, \dots, j_2 - 1$, $p_{kl} = 0$ for all $l = 0, \dots, k$.

(C2) For all $k = j_2 + 1, \dots, J - 1$, $p_{kl} = 0$ for all $l = j_2 + 1, \dots, k$.

Similarly, because the second equality in (23) is equivalent to

$$\sum_{k \geq j_2} \sum_{l > j_2} p_{kl} = \sum_{k \geq j_2} \sum_{l > j_2} p_{kl} + \sum_{k < j_2} \sum_{l > j_2} p_{kl},$$

the probability matrix \mathbf{P} must further satisfy the following constraint:

(C3) $p_{kl} = 0$, for all $k = 0, \dots, j_2 - 1$ and $l = j_2 + 1, \dots, J - 1$.

The constraints in (C1), (C2) and (C3) imply that \mathbf{P} must have the following block structure:

$$\mathbf{P} = \begin{pmatrix} (\mathbf{0}, \mathbf{P}_{\text{tl}}) & \mathbf{0} \\ \mathbf{P}_{\text{bl}} & \begin{pmatrix} \mathbf{P}_{\text{br}} \\ \mathbf{0}^T \end{pmatrix} \end{pmatrix} \quad (25)$$

where the $j_2 \times j_2$ sub-matrix \mathbf{P}_{tl} and the $(J - j_1 - 1) \times (J - j_1 - 1)$ sub-matrix \mathbf{P}_{br} are both upper triangular, and the $(J - j_2) \times (j_2 + 1)$ sub-matrix \mathbf{P}_{bl} on bottom left has no restrictions.

Because $p_{+j_2} + \Delta_{j_2} \geq p_{+j} + \Delta_j$ for all j , we have

$$\sum_{k=j}^{j_2-1} p_{k+} \leq \sum_{l=j}^{j_2-1} p_{+,l+1} \quad (j = 0, \dots, j_2 - 1); \quad \sum_{k=j_2}^s p_{k+} \geq \sum_{l=j_2}^s p_{+,l+1} \quad (j = j_2, \dots, J - 2).$$

Given the above two sets of constraints for the marginal probabilities, we construct the probability matrix \mathbf{P} in three steps.

- (1) We apply Lemma 1(b) to $(p_{0+}, \dots, p_{j_2-1,+})$ and $(p_{+1}, \dots, p_{+,j_2})$, and obtain an upper triangular matrix $\mathbf{P}_{\text{tl}} = (p_{kl})_{0 \leq k \leq j_2-1, 1 \leq l \leq j_2}$ with nonnegative elements such that

$$\sum_{l'=1}^{j_2} p_{kl'} = p_{k+}, \quad \sum_{k'=0}^{j_2-1} p_{k'l} \leq p_{+l} \quad (k = 0, \dots, j_2 - 1; l = 1, \dots, j_2).$$

- (2) We apply Lemma 1(d) to $(p_{j_2+}, \dots, p_{J-2,+})$ and $(p_{+,j_2+1}, \dots, p_{+,J-1})$, and obtain an upper triangular matrix $\mathbf{P}_{\text{br}} = (p_{kl})_{j_2 \leq k \leq J-2, j_2+1 \leq l \leq J-1}$ with nonnegative elements such that

$$\sum_{l'=j_2+1}^{J-1} p_{kl'} \leq p_{k+}, \quad \sum_{k'=j_2}^{J-2} p_{k'l} = p_{+l} \quad (k = j_2, \dots, J-2; l = j_2+1, \dots, J-1).$$

- (3) We construct $\mathbf{P}_{\text{bl}} = (p_{kl})_{j_2 \leq k \leq J-1, 0 \leq l \leq j_2}$ by letting

$$p_{kl} = \left(p_{k+} - \sum_{l'=j_2+1}^{J-1} p_{kl'} \right) \left(p_{+l} - \sum_{k'=0}^{j_2-1} p_{k'l} \right) \geq 0 \quad (k = j_2, \dots, J-1; l = 0, \dots, j_2).$$

The constructed probability matrix \mathbf{P} has marginal probabilities $\mathbf{p}_1 = (p_{0+}, \dots, p_{J-1,+})^T$ and $\mathbf{p}_0 = (p_{+0}, \dots, p_{+,J-1})^T$. Moreover, by (25) the corresponding τ is the sum of all the elements in \mathbf{P}_{bl} , which we construct in the above (3). Therefore,

$$\tau = 1 - \sum_{k'=0}^{j_2-1} p_{k'+} - \sum_{l'=j_2+1}^{J-1} p_{+l'} = \sum_{k'=j_2}^{J-1} p_{k'+} - \sum_{l'=j_2+1}^{J-1} p_{+l'} = p_{+j_2} + \Delta_{j_2},$$

which implies that \mathbf{P} attains τ_L . □

References

- Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd ed.* Hoboken, New Jersey: John Wiley and Sons.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Assoc.*, 91:444–455.

- Baker, S. G. (2011). Estimation and inference for the causal effect of receiving treatment on a multinomial outcome: An alternative approach. *Biometrics*, 67:319–323.
- Beran, R. (1988). Balanced simultaneous confidence sets. *J. Am. Statist. Assoc.*, 83:679–697.
- Beran, R. (1990). Refining bootstrap simultaneous confidence sets. *J. Am. Statist. Assoc.*, 85:517–426.
- Bradley, R. A., Katti, S. K., and Coons, I. J. (1962). Optimal scaling for ordered categories. *Psychometrika*, 27:355–374.
- Cheng, J. (2009). Estimation and inference for the causal effect of receiving treatment on a multinomial outcome. *Biometrics*, 65:96–103.
- Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *J. R. Statist. Soc. B*, 68:815–836.
- Chernozhukov, V., Lee, S., and Rosen, A. (2013). Intersection bounds: estimation and inference. *Econometrica*, 81:667–737.
- Demidenko, E. (2016). The p-value you can’t buy. *Am. Statistician*, in press.
- Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39:1–38.
- Díaz, I., Colantuoni, E., and Rosenblum, M. (2016). Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics*, 72:422–431.
- Ding, P. and Dasgupta, T. (2016). A potential tale of two by two tables from completely randomized experiments. *J. Am. Statist. Assoc.*, 111:157–168.
- Djebbari, H. and Smith, J. A. (2008). Heterogeneous impacts in PROGRESA. *J. Econometrics*, 145:64–80.
- Fan, Y. and Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Economet. Theor.*, 26:931–951.

- Fan, Y., Sherman, R., and Shum, M. (2014). Identifying treatment effects under data combination. *Econometrica*, 82:811–822.
- Frank, M. J., Nelsen, R. B., and Schweizer, B. (1987). Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probab. Theor. Rel.*, 74:199–211.
- Freedman, D. A. (2008). Randomization does not justify logistic regression. *Statis. Sci.*, 23:237–249.
- Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Am. Statist. Assoc.*, 107:450–466.
- Gadbury, G. L. and Iyer, H. K. (2000). Unit-treatment interaction and its practical consequences. *Biometrics*, 56:882–885.
- Grilli, L. and Mealli, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *J. Educ. Behav. Stat.*, 33:111–130.
- Heckman, J. J., Smith, J., and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.*, 64:487–535.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189.
- Hirano, K. and Porter, J. (2012). Impossibility results for nondifferentiable functionals. *Econometrica*, 80:1769–1790.
- Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Statist. Assoc.*, 95:77–84.
- Huang, E., Fang, E., Hanley, D., and Rosenblum, M. (2015). Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few. Working Paper 274, Johns Hopkins University, Department of Biostatistics.
- Ju, C. and Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *J. R. Statist. Soc. B*, 72:129–142.

- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev. Econ. Stud.*, 76:1071–1102.
- Long, D. M. and Hudgens, M. G. (2013). Sharpening bounds on principal effects with covariates. *Biometrics*, 69:812–819.
- Makarov, G. D. (1982). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory Probab. Appl.*, 26:803–806.
- Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Am. Statist. Assoc.*, 108:1120–1131.
- Nelsen, R. B. (2006). *An Introduction to Copulas, Second Edition*. New York: Springer.
- Newcombe, R. G. (2006a). Confidence intervals for an effect size measure based on the mann–whitney statistic. part 1: general issues and tail-area-based methods. *Stat. Med.*, 25:543–557.
- Newcombe, R. G. (2006b). Confidence intervals for an effect size measure based on the mann–whitney statistic. part 2: asymptotic methods and evaluation. *Stat. Med.*, 25:559–573.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. Section 9. *Statist. Sci.*, 5:465–472.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference, Second Edition*. Cambridge University Press.
- Richardson, A., Hudgens, M. G., Gilbert, P. B., and Fine, J. P. (2014). Nonparametric bounds and sensitivity analysis of treatment effects. *Stat. Sci.*, 29:596–618.
- Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *J. Stat. Plan. Inference*, 138:2786–2807.
- Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, 78:169–211.
- Rosenbaum, P. R. (2001). Effects attributable to treatment: inference in experiments and observational studies within a discrete pivot. *Biometrika*, 88:219–231.

- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66:688–701.
- Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Stat.*, 6:34–58.
- Rubin, D. B. (1980). Comment on “Randomization analysis of experimental data: the Fisher randomization test” by D. Basu. *J. Am. Statist. Assoc.*, 75:591–593.
- Rüschendorf, L. (1982). Random variables with maximum sums. *Adv. in Appl. Probab.*, 14:623–632.
- Scharfstein, D. O., Manski, C. F., and Anthony, J. C. (2004). On the construction of bounds in prospective studies with missing ordinal outcomes: application to the good behavior game trial. *Biometrics*, 60:154–164.
- Schochet, P. Z., Cao, J. B. R., Glazerman, S., Grady, A., Gritz, M., McConnell, S., Johnson, T., and Burghardt, J. (2003). National job corps study: Data documentation and public use files, Volume I. *Documentation, Washington, DC, Mathematica Policy Research, Inc.*
- Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Stat.*, 36:423–439.
- Volfovsky, A., Airoidi, E. M., and Rubin, D. B. (2015). Causal inference for ordinal outcomes. *arXiv*, 1501.01234.
- Yang, F. (2014). Causal inference methods for addressing censoring by death and unmeasured confounding using instrumental variables. *Ph.D. Thesis, University of Pennsylvania.*
- Yang, F. and Small, D. S. (2016). Using post-quality of life measurement information in censoring by death problems. *J. R. Statist. Soc. B*, 78:299–318.
- Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *J. Am. Statist. Assoc.*, 104:166–176.
- Zhou, W. (2008). Statistical inference for $P(X < Y)$. *Stat. Med.*, 27:257–279.

Supplementary Material

The Supplementary Material consists of two parts. In Part A, we prove Lemma 1 introduced in the Appendix, and provide the proofs of all the theorems and corollaries in the main text, except for Theorem 1. In Part B, we present the sufficient and necessary conditions for the bounds in Theorem 1 to be the same.

A. Proof of Lemma, Theorems and Corollaries

A.1. Proof of Lemma 1

Proof of Lemma 1(a). We prove by induction. When $n = 1$, we let $\mathbf{A}_1 = y_0 \geq 0$, and Lemma 1(a) holds because $y_0 \leq x_0$. When $n \geq 2$, suppose Lemma 1(a) holds for $n - 1$. In particular, for any (x_1, \dots, x_{n-1}) and (y_1, \dots, y_{n-1}) such that $\sum_{r=s}^{n-1} x_r \geq \sum_{r=s}^{n-1} y_r$ for all $s = 1, \dots, n - 1$, there exists a lower triangular matrix $\mathbf{A}_{n-1} = (a_{kl})_{1 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=1}^{n-1} a_{kl'} \leq x_k, \quad \sum_{k'=1}^{n-1} a_{k'l} = y_l \quad (k, l = 1, \dots, n - 1). \quad (\text{A.1})$$

To prove that Lemma 1(a) holds for n , we let

$$\mathbf{A}_n = \begin{pmatrix} a_{00} & \mathbf{0}^T \\ \mathbf{a} & \mathbf{A}_{n-1} \end{pmatrix},$$

where a_{00} and $\mathbf{a} = (a_{10}, \dots, a_{n-1,0})^T$ are defined for two separate cases below.

- (1) $y_0 < x_0$. We let $a_{00} = y_0$, and $a_{k0} = 0$ for all $k = 1, \dots, n - 1$. Clearly, \mathbf{A}_n has nonnegative elements, and satisfies the row and column sum conditions in Lemma 1(a) holds;
- (2) $y_0 \geq x_0$. We let $a_{00} = x_0$, and

$$a_{k0} = (y_0 - a_{00}) \frac{x_k - \sum_{l'=1}^{n-1} a_{kl'}}{\sum_{k'=1}^{n-1} (x_{k'} - \sum_{l'=1}^{n-1} a_{k'l'})} \geq 0 \quad (k = 1, \dots, n - 1). \quad (\text{A.2})$$

This construction guarantees that the column sums of \mathbf{A}_n are y_l 's. Furthermore, because

\mathbf{A}_{n-1} satisfies (A.1), we have

$$\begin{aligned} \sum_{k'=1}^{n-1} \left(x_{k'} - \sum_{l'=1}^{n-1} a_{k'l'} \right) &= \sum_{k'=1}^{n-1} x_{k'} - \sum_{k'=1}^{n-1} \sum_{l'=1}^{n-1} a_{k'l'} = \sum_{k'=1}^{n-1} x_{k'} - \sum_{l'=1}^{n-1} \sum_{k'=1}^{n-1} a_{k'l'} \\ &= \sum_{k'=1}^{n-1} x_{k'} - \sum_{k'=1}^{n-1} y_{k'} \geq y_0 - x_0 = y_0 - a_{00} > 0. \end{aligned} \quad (\text{A.3})$$

Formulas (A.2) and (A.3) imply that $a_{k0} \leq x_k - \sum_{l'=1}^{n-1} a_{kl'}$ and therefore $\sum_{l'=0}^{n-1} a_{kl'} \leq x_k$ for $k = 1, \dots, n-1$.

Therefore Lemma 1(a) holds for n , and the proof is complete. \square

Proof of Lemma 1(b). By applying Lemma 1(a) to (y_0, \dots, y_{n-1}) and (x_0, \dots, x_{n-1}) , we obtain a lower triangular matrix $\widetilde{\mathbf{B}}_{\mathbf{n}} = (\tilde{b}_{kl})_{0 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{k'=0}^{n-1} \tilde{b}_{k'l} = x_k, \quad \sum_{l'=0}^{n-1} \tilde{b}_{kl'} \leq y_k \quad (k, l = 0, \dots, n-1).$$

Let $\mathbf{B}_n = \widetilde{\mathbf{B}}_{\mathbf{n}}^T$, and the proof is complete. \square

Proof of Lemma 1(c). By applying Lemma 1(a) to (y_{n-1}, \dots, y_0) and (x_{n-1}, \dots, x_0) , we obtain a lower triangular matrix $\widetilde{\mathbf{C}}_{\mathbf{n}} = (\tilde{c}_{kl})_{0 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{k'=0}^{n-1} \tilde{c}_{k'l} = x_{n-l-1}, \quad \sum_{l'=0}^{n-1} \tilde{c}_{kl'} \leq y_{n-k-1} \quad (k, l = 0, \dots, n-1).$$

Let $\mathbf{C}_n = (\tilde{c}_{n-l-1, n-k-1})_{0 \leq k, l \leq n-1}$, and the proof is complete. \square

Proof of Lemma 1(d). By applying Lemma 1(c) to (y_0, \dots, y_{n-1}) and (x_0, \dots, x_{n-1}) , we obtain a lower triangular matrix $\widetilde{\mathbf{D}}_{\mathbf{n}} = (\tilde{d}_{kl})_{0 \leq k, l \leq n-1}$ with nonnegative elements such that

$$\sum_{l'=0}^k \tilde{d}_{kl'} = y_k, \quad \sum_{k'=l}^{n-1} \tilde{d}_{k'l} \leq x_k \quad (k, l = 0, \dots, n-1).$$

Let $\mathbf{D}_n = \widetilde{\mathbf{D}}_{\mathbf{n}}^T$, and the proof is complete. \square

Proof of Lemma 1(e). In addition to the proof of Lemma 1(a), we further need to show that if

$\sum_{r=0}^{n-1} y_r = \sum_{r=0}^{n-1} x_r$, the row sums of the constructed matrix \mathbf{A}_n are x_k 's. In the induction of the proof of Lemma 1(a), if we have constructed matrix A_{n-1} , the case with $y_0 < x_0$ would not happen. We consider only the case with $y_0 \geq x_0$. Because the lower triangular matrix A_{n-1} has the column sums y_l 's, and $\sum_{r=0}^{n-1} y_r = \sum_{r=0}^{n-1} x_r$, we have

$$\sum_{k'=1}^{n-1} \left(x_{k'} - \sum_{l'=1}^{n-1} a_{k'l'} \right) = \sum_{k'=1}^{n-1} x_{k'} - \sum_{k'=1}^{n-1} y_{k'} = y_0 - x_0 = y_0 - a_{00} > 0.$$

The above formula, coupled with the construction of the first column of \mathbf{A}_n in (A.2), gives $a_{k0} = x_k - \sum_{l'=1}^{n-1} a_{kl'}$ and thus $\sum_{l'=0}^{n-1} a_{kl'} = x_k$ for all k . \square

A.2. Proofs of Other Theorems and Corollaries

Proof of Theorem 2. Because $\eta = 1 - \text{pr}\{Y_i(0) \geq Y_i(1)\}$, its lower bound is one minus the upper bound of $\text{pr}\{Y_i(0) \geq Y_i(1)\}$. By switching the treatment and control labels, we can bound $\text{pr}\{Y_i(0) \geq Y_i(1)\}$ from the above by

$$\text{pr}\{Y_i(0) \geq Y_i(1)\} \leq 1 - \max_{0 \leq j \leq J-1} \Delta_j,$$

which implies that $\eta_L = \max_{0 \leq j \leq J-1} \Delta_j$.

Similarly, the upper bound of η equals one minus the lower bound of $\text{pr}\{Y_i(0) \geq Y_i(1)\}$. By switching the treatment and control labels, we can bound $\text{pr}\{Y_i(0) \geq Y_i(1)\}$ from below by

$$\text{pr}\{Y_i(0) \geq Y_i(1)\} \geq \max_{0 \leq j \leq J-1} (p_{j+} - \Delta_j),$$

which implies that $\eta_U = 1 + \min_{0 \leq j \leq J-1} (\Delta_j - p_{j+})$. \square

Proof of Theorem 3. With independent potential outcomes, the probability matrix P has elements $p_{kl} = p_{k+}p_{+l}$ for k and l . We obtain τ_I and η_I by their definitions. Obviously, they are between their lower and upper bounds, i.e., $\tau_L \leq \tau_I \leq \tau_U$ and $\eta_L \leq \eta_I \leq \eta_U$. \square

Proof of Theorem 4. The proof follows Lee (2009). Because any value of τ within the covariate adjusted bounds $[\tau'_L, \tau'_U]$ must be compatible with the distributions of $\{Y(1), \mathbf{X}\}$ and $\{Y(0), \mathbf{X}\}$, it must also be compatible with the distributions of $Y(1)$ and $Y(0)$ by discarding \mathbf{X} . Therefore, any

value of τ within the adjusted bounds $[\tau'_L, \tau'_U]$ must also be within the unadjusted bounds $[\tau_L, \tau_U]$. Consequently, the adjusted bounds are tighter, i.e., $[\tau'_L, \tau'_U] \subset [\tau_L, \tau_U]$. Similar arguments apply to the covariate adjusted bounds and the unadjusted bounds for τ_c . \square

Proof of Theorem 5. Under monotonicity, by the law of total probability, we have

$$\tau = \pi_c \tau_c + \pi_a \tau_a + \pi_n \tau_n.$$

Under exclusion restriction, we have $\tau_a = 1$ and $\tau_n = 1$, yielding

$$\tau = \pi_c \tau_c + 1 - \pi_c,$$

which implies that

$$\tau_c = \tau / \pi_c - (1 - \pi_c) / \pi_c.$$

Analogously, we have $\eta = \pi_c \eta_c$, which implies that $\eta_c = \eta / \pi_c$. \square

Proof of Corollary 1. By Theorem 1, $\tau = 1$ if and only if $\min_{0 \leq j \leq J-1} \Delta_j = 0$. Because $\Delta_0 = 0$, this is equivalent to $\Delta_j \geq 0$ for all j , i.e., the stochastic dominance assumption holds. \square

Proof of Corollary 2. The proof follows directly from Theorems 1 and 2. \square

Proof of Corollary 3. The closed-form expressions for $\tau''_{c,L}$, $\tau''_{c,U}$, $\eta''_{c,L}$ and $\eta''_{c,U}$ follow directly from Theorem 5 and Corollary 2. Furthermore, under the monotonicity and exclusion restriction assumptions, we have

$$\Delta_j = \pi_c \Delta_{c,j} \quad (j = 0, \dots, J-1).$$

Therefore, for the upper bound of τ , we have

$$\tau_U = 1 - \pi_c + \pi_c(1 + \min \Delta_{c,j}) = \tau''_U,$$

and for the lower bound, we have

$$\tau_L \leq \max(p_{+j} - 1 + \pi_c + \pi_c \Delta_{c,j}) = 1 - \pi_c + \pi_c \max(c_{+j} + \Delta_{c,j}) = \tau''_L.$$

The first step holds because under the strong monotonicity assumption $n_{+j}\pi_n \leq \pi_n$, and under the monotonicity assumption $a_{+j}\pi_a + n_{+j}\pi_n \leq \pi_a + \pi_n$. Similar arguments apply to the bounds of η . \square

B. Condition for Point Identification of τ and η

Under certain condition, the lower and upper bounds of τ (or η) in Theorem 1 will be the same, resulting in identification of the corresponding causal parameters. We formally state the sufficient and necessary conditions for this to happen in the following theorem.

Theorem B.1. Let $\mathbb{K} = \{k : p_{k+} > 0\}$ and $\mathbb{L} = \{l : p_{+l} > 0\}$. The lower and upper bounds of τ are the same, if and only if there does not exist $k_1, k_2 \in \mathbb{K}$ and $l_1, l_2 \in \mathbb{L}$ such that

$$k_2 \geq l_2 > k_1 \geq l_1 \quad \text{or} \quad l_2 > k_2 \geq l_1 > k_1. \quad (\text{B.1})$$

The lower and upper bounds of η are the same, if and only if there does not exist $k_1, k_2 \in \mathbb{K}$ and $l_1, l_2 \in \mathbb{L}$ such that

$$l_2 \geq k_2 > l_1 \geq k_1 \quad \text{or} \quad k_2 > l_2 \geq k_1 > l_1. \quad (\text{B.2})$$

Proof. Similar to the proof of Theorem 2, because $\eta = 1 - \text{pr}\{Y_i(0) \geq Y_i(1)\}$, (B.1) immediately implies (B.2). Therefore, we need only to prove that (B.1) is the sufficient and necessary condition that the lower and upper bounds of τ are the same, i.e., $\tau_L = \tau_U$.

First we prove the necessity of the condition. Assume that it does not hold, i.e., there does exist $k_1, k_2 \in \mathbb{K}$ and $l_1, l_2 \in \mathbb{L}$ such that (B.1) holds. In this case we construct two probability matrices with the same marginal probabilities but different values of τ . The first probability matrix is $\mathbf{P} = (p_{k+p+l})_{0 \leq k, l \leq J-1}$. For the second probability matrix, let $\xi = \min(p_{k_1, +p+, l_1}, p_{k_2, +p+, l_2})$, which is a positive constant. We then apply the following matrix operation to the 2×2 sub-matrix of the first probability matrix:

$$\begin{pmatrix} p_{k_1 l_1} & p_{k_1 l_2} \\ p_{k_2 l_1} & p_{k_2 l_2} \end{pmatrix} \longrightarrow \begin{pmatrix} p_{k_1 l_1} - \xi & p_{k_1 l_2} + \xi \\ p_{k_2 l_1} + \xi & p_{k_2 l_2} - \xi \end{pmatrix}$$

The above operation preserves the marginal probabilities, and the difference of τ between the first

and second probability matrices is ξ , if $k_2 \geq l_2 > k_1 \geq l_1$, and $-\xi$, if $l_2 > k_2 \geq l_1 > k_1$.

Second, we prove the sufficiency of the condition. If $|\mathbb{K}| = 1$ or $|\mathbb{L}| = 1$, the probability matrix degenerates and consequently we have $\tau_L = \tau_{c,U}$. If $|\mathbb{K}| \geq 2$ and $|\mathbb{L}| \geq 2$, let $k_* = \min_{k \in \mathbb{K}} k$ and $k^* = \max_{k \in \mathbb{K}} k$ be the minimal and maximal indices of nonzero p_{k+} 's, and $l_* = \min_{l \in \mathbb{L}} l$ and $l^* = \max_{l \in \mathbb{L}} l$ the minimal and maximal indices of nonzero p_{+l} 's. A useful fact that we repeatedly use is that if $p_{k+} = 0$, then $p_{kl} = 0$ for all l . Similarly, if $p_{+l} = 0$, then $p_{kl} = 0$ for all k .

Because k_*, k^* and l_*, l^* cannot satisfy (B.1), we discuss the two following cases based on the relative locations of the two intervals $[k_*, k^*]$ and $[l_*, l^*]$:

1. “Non-overlapping,” i.e., $k_* \geq l^*$ or $k^* < l_*$:

(a) If $k_* \geq l^*$, we prove that $p_{kl} = 0$ for all $k < l$. Assume the claim does not hold, then there exists $k' < l'$ such that $p_{k'l'} > 0$, then $p_{k'+} > 0$ and $p_{+l'} > 0$. This implies that $k_* \leq k' < l' \leq l^*$, contradicting the initial assumption. Therefore, $\tau_L = \tau_U = 1$.

(b) If $k^* < l_*$, similarly $p_{kl} = 0$ for all $k \geq l$, implying that $\tau_L = \tau_U = 0$.

2. “Inclusive,” i.e., $l^* > k^* > k_* \geq l_*$ or $k^* \geq l^* > l_* > k_*$:

(a) If $l^* > k^* > k_* \geq l_*$, and furthermore if there exists $l' \in \mathbb{L}$ such that $k_* < l' \leq k^*$, then $l' \neq l_*$ and $l' \neq l^*$. Moreover, k_*, k^* and l', l^* satisfy (B.1), contradicting the initial assumption. Therefore for all $l \in \mathbb{L}$, $l \leq k_*$ or $l > k^*$. Consequently,

$$\begin{aligned} \tau &= \sum_{k \geq l} p_{kl} 1(k \in \mathbb{K}, l \in \mathbb{L}) = \sum_{k \geq l} p_{kl} 1(k \in \mathbb{K}, l \in \mathbb{L}, l \leq k_*) \\ &= \sum p_{kl} 1(k \in \mathbb{K}, l \in \mathbb{L}, l \leq k_*) = \sum_{l \leq k_*, l \in \mathbb{L}} p_{+l} \end{aligned}$$

is identifiable, which implies that $\tau_L = \tau_U$.

(b) If $k^* \geq l^* > l_* > k_*$, similarly as above for all $k \in \mathbb{K}$, $k < l_*$ or $k \geq l^*$. Consequently,

$$\tau = \sum_{k \geq l} p_{kl} 1(k \in \mathbb{K}, l \in \mathbb{L}) = \sum_{k \geq l} p_{kl} 1(k \in \mathbb{K}, l \in \mathbb{L}, k \geq l^*) = \sum_{k \geq l^*, k \in \mathbb{K}} p_{k+}$$

is identifiable, which implies that $\tau_L = \tau_U$.

